基于遗传算法的空间网格划分匿名算法 *

孙 悦[a]，张 磊[a+]，李 晶[a]，张 震[b]

(佳木斯大学 a. 信息电子技术学院; b. 机械工程学院，黑龙江 佳木斯 154007)

摘 要：隐私泄露问题已经成为阻碍基于位置的服务(location-based services，LBS)进一步发展的原因。针对当 LBS 用户发送查询时，用户的个人隐私可能会泄露给攻击者的问题，提出了基于遗传算法的空间网格划分的隐私保护算法（简称 GAGP）。算法包括两个方法，地图分割算法和假名生成法。地图分割算法利用遗传算法给每个网格赋权值，再通过使用邻接网格扩展的方法，保证每个划分区域的查询频率基本相等。假名生成法是用户在每次发送查询时使用假名来应对长期统计的攻击方式。通过实验证明所提算法与其他三种算法相比结果较好，所以提出的方案能够有效的保护用户的隐私。

关键词：位置隐私保护；网格划分；假名；遗传算法；位置服务

中图分类号：TP309.2　　doi: 10.19734/j.issn.1001-3695.2018.09.0753

Anonymous algorithm for spatial mesh generation based on genetic algorithm

Sun Yue[a], Zhang Lei[a+], Li Jing[a], Zhang Zhen[b]

(a. *School of Information & Electronic Technology*, *b. School of Mechanical Engineering*, *Jiamusi University*, *Jiamusi Heilongjiang* 154007, *China*)

Abstract: Privacy breaches have become an obstacle to the further development of location-based services (location-based services, LBS) . Concerns that when a LBS user sends a query, the user's personal privacy may be disclosed to an attacker. This paper proposed a scheme called grid-based genetic privacy protection algorithm (short for GAGP) that based on the conception of weighted optimal genetic algorithm. This scheme involved two basic procedures: map segmentation and pseudonym generation. Map segmentation algorithm uses genetic algorithm to assign values to each grid, and then uses the method of adjacent grid expansion to ensure that the query frequency of each partition area is basically equal. Kana generation is a way for users to use a pseudonym each time they send a query in response to a long-term statistical attack. The experimental results show that the proposed algorithm is better than the other three algorithms, so the proposed scheme can effectively protect the privacy of users.

Key words: location privacy protection; grid generation; pseudonym; genetic algorithm; location-based service

0 引言

当今社会网络定位技术不断发展，位置服务(location based service，LBS) 已经是人们生活中不可或缺的服务[1,2]。随着 LBS 给用户提供服务的，同时也给人们带来了位置隐私泄露的重要问题，因此能否解决用户隐私保护的需要也就成为了公众安心使用 LBS 服务的重要前提[3~5]。随着技术的逐步发展，出现了多种对位置数据隐私保护的方法[6~8]。例如，假位置[9]、位置 k-匿名[10]、空间加密[11-12]等方法，其中使用最为广泛的是 k-匿名方法[13~16]。其中最开始使用的 Random 不考虑查询概率，通过随机选择生成匿名区域的算法；以及后来的 GridDummy[17]是和在满足用户隐私要求的情况下获取匿名区域的方法；再到 en-DLS[18] 算法是基于从用户的历史位置提出的虚拟位置选择的算法。但都仅通过用户的位置信息来保护用户位置隐私，不能应对针对长期统计的攻击方法以达到隐私保护的目的。

为了解决上述讨论的问题，本文提出 GAGP 算法。首先用户可以进行网格的预划分，代理可以根据所获得的历史查询数据计算出每个网格提交查询的概率，使用基于遗传算法的方法算出权值。然后根据水平扩展方法来扩张网格区域，保证每个查询区域的权值基本相同即每个匿名区域的查询频率基本相同，最后采用多假名的方法防止 LBS 服务器的长期统计攻击。实验中使用真实数据对 GAGP 算法进行实验仿真与分析，并与其他三种算法进行对比，最终实验结果验证了 GAGP 算法的良好性能。

1 准备工作

1.1 LBS 服务

在现有的 LBS 查询方法中，用户向 LBS 服务器提交查询前，移动用户首先通过内置于智能手机的 GPS/WIFI 模块来获取自己的当前位置。然后智能手机直接或者通过第三方服务器将查询内容发送到 LBS 服务器，包括标志符，确切位置，兴趣和查询范围等。最后 LBS 服务器将根据用户的查询反馈 POI，如图 1 所示。

为了使用户的隐私信息不受到侵害，在传统的隐私保护算法的方法中，选择 k-1 个用户建立协作组来混淆攻击者来

收稿日期：2018-09-16；**修回日期：**2018-10-22　**基金项目：**黑龙江省普通本科高等学校青年创新人才培养计划资助项目（UNPYSCT-2017149）；国家级大学生创新创业训练项目（201810222033）

作者简介：孙悦（1995-），吉林梨树人，硕士，主要研究方向为隐私保护、数据挖掘；张磊（1982-），男（通信作者），讲师，博士，主要研究方向为隐私保护、信息安全（1458516851@qq.com）；李晶（1968-），女，教授，硕士，主要研究方向为网络安全、数据挖掘；张震（1994-），男，硕士，主要研究方向为隐私保护、人工智能.

保护用户的真实位置，但协作用户的位置常常很难选择，在较为密集的地方组成的 *k*-匿名集合的用户通常非常接近真实用户导致匿名区域较小。使用虚拟位置可以用来解决这些问题，然而大多数现有作品一般使用随机的方法来选择虚拟位置。如下例所示，这并不是一个最好的解决方法。例如 Alice 发送查询，在随机方案中，虚拟位置是随机选择的，这是 Alice 会认为自己被攻击者发现的概率是，这也是 k-匿名的理论结果，但由于攻击者会根据位置查询概率的辅助信息过滤掉一些过分虚假的位置，使得隐私级别降低，如图 2 所示。

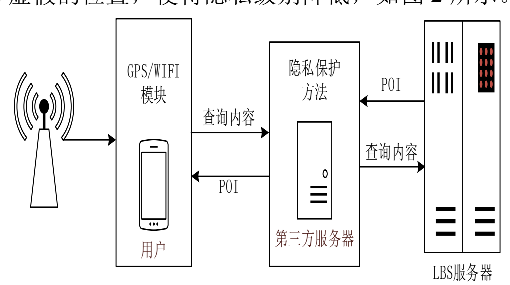


图 1　LBS 的系统结构

Fig. 1　The system structure of LBS

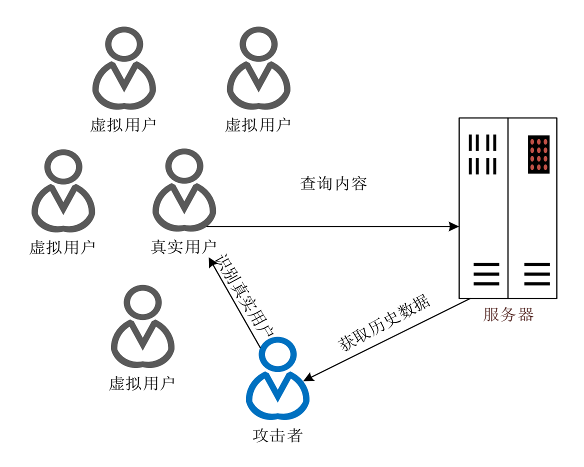


图 2　通过分析虚假位置进行攻击的方法

Fig. 2　Method of attack by analyzing false position

攻击者的目标是想要获取用户的隐私信息，其中包括了姓名、兴趣点和用户真实位置。攻击者可以监视某个区域，以获取用户发出的隐私信息。攻击者也可以以病毒的形式侵入到 LBS 服务器，查看 LBS 服务器发送给用户的 POI，从而推断出用户身份和位置等信息。攻击者还可以直接在侵入的 LBS 服务器中获取到历史数据，因此在本文中，以 LBS 服务器为攻击者。出于利益，他试图获取与用户相关的隐私信息，并且对用户的真实位置以及 LBS 查询感兴趣。他能够获取系统中的所有信息，不仅知道用户当前的 LBS 查询，还可以获取用户的历史数据。

1.2　研究动机和基本思想

由于攻击者可掌握某一指定用户的背景知识（如用户的查询频率），并利用这些知识作为辅助信息来识别用户的真实位置。针对这样一种潜在的攻击行为，本文的基础思想是利用优化的网格划分来应对长期统计攻击和区域攻击。基于这样一种思想提出了 GAGP 的隐私保护方法。该方法首先由用户确定好初始的网格划分程度，并通过遗传算法获取到每个单元格的权值，将这些权值求和取平均得到阈值；其次使用邻接网格扩展匿名算法进行网格的扩展，即从第一个网格开始，与水平临接网格的权值进行相加，并判断当前权值之和是否等于平均权值，若相等则返回该网格区域。若不相等则根据权值之和与平均权值的差是否满足相应条件来选择是否继续扩展，该方法以递归的形式对未进入匿名区域的单元格执行上一步算法直至所有的单元网格遍历结束。最后为针对长期统计攻击，本文还使用更换假名的方式，将两种方式结合在一起，从而保证用户得真实位置不受侵害。

2　GAGP 算法

本文算法旨在保护用户的位置隐私，使在攻击者掌握历史查询频率的情况下，仍能使用户查询位置所处的匿名空间与其他匿名空间的查询频率相同。基于该思想，本文的隐私保护方法需使用第三方作为代理，首先用户对空间网格进行预划分，代理可以根据历史查询数据计算出每个网格提交查询的概率，使用基于遗传算法的方法算出权值。然后根据水平扩展方法来扩张网格区域以保证每个匿名区域的查询概率基本相同。最后还采用了多假名的方法以应对 LBS 服务器长期统计攻击。具体处理过程可参照图 3 所示。

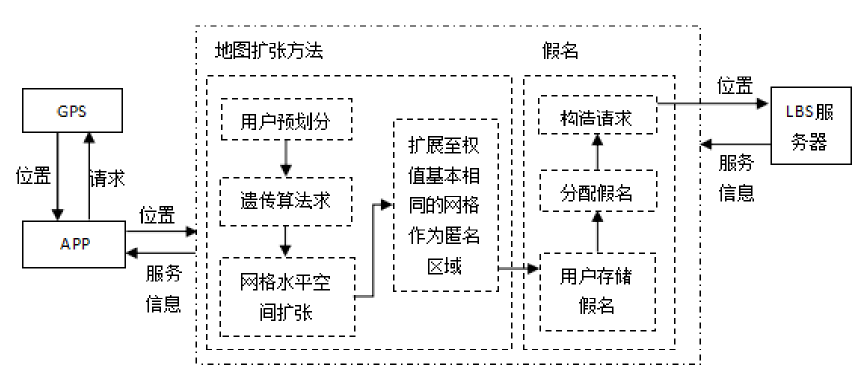


图 3　GAGP 算法方案图

Fig. 3　GAGP algorithm scheme diagram

2.1　网格预划分

用户在第三方程序上注册登录并授予权限，并且假设代理可以获取整个地区的历史数据的情况。用户根据自己隐私保护的需求，设置自己的网格划分程度。这里设置为 n，即用户设置 n 后，将整个区域分成 $n \times n$ 的网格，代理可以根据历史查询数据计算出每个网格提交查询的概率，如图 4 所示。其中颜色较深的为查询频率较高的地区，颜色越浅说明该区域查询频率越低。

2.2　网格的扩张方法

本文提出的匿名算法首先采用遗传算法对用户划分的 $n \times n$ 网格进行赋权值，累加权值求和再求平均值，以此平均值为阈值，采用水平空间扩展的方法。从第一个单元格开始，向水平方向临近单元格扩展，判断累加的权值是否等于阈值，若相等则返回该网格区域，若不等则继续扩展。该方法采用

递归的形式对未进入匿名区域的单元格执行此前的算法，直至所有的单元网格都遍历结束。算法 1 描述了区域扩展的过程。

算法 1　区域扩张算法

输入：各个单元格 a_{ij} 的权值 w_{ij}，以及平均值 avg，权值累加和 sum=0。

输出：被组合成同一匿名区域的单元格的集合。

```
for(i=0; i<n; i++)
  for(j=0; j<n; j=+)
    if(sum<avg)
      sum=sum+avg;
      并将当前 aij 存到集合 n 中；
    end if
    比较现在这个 sum 值和上一个 sum 值与 avg 的差的绝对值哪个更小；
    if 当前 sum 值更合就把当前的 aij 也存到 n 中，保存集合 n；
    else 就返回上一个 aij，保存集合 n；
return n;
```

网格匿名空间的划分的典型方法有四分网格法及四叉树划分网格算法，四分网格算法与四叉树划分网格的方法相比有形成匿名区域小、精度高的优点。而本文算法是先计算网格的查询频率并赋权值，根据权值来向周围网格进行扩展，与四分网格算法相比，粒度和产生的匿名空间都要更小一些。且在后续算法的优化中使用了假名，通过假名来针对基于长期统计的攻击，用户使用不同的假名来迷惑攻击者，使其无法确认真实用户的隐私信息。

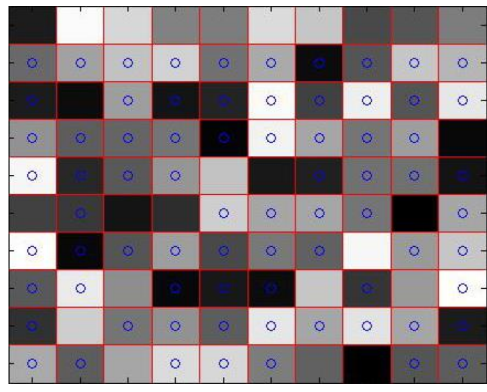


图 4　网格预划分结果

Fig. 4　Grid predivision result

2.3　算法的优化

由于地形和居住地点布局的不同，在各地区应用的 LBS 服务的概率也各不相同，比如用户一般不会在有山川河流的这些地区申请 LBS 服务。因此，选择 k-1 个虚拟位置的传统方法由于边信息就不能有效地保护真实位置。针对算法 1 的区域扩张算法来说，只能使划分后的的查询频率基本相同。但一旦存在长期统计攻击，攻击者仍可以判断出用户的真实位置。因此，本文采取多假名的方式来应对这种类型的攻击。初始时为用户请求分配一个假名，且用户存储多个用户假名，每次应用 LBS 服务时，用户便从诸多假名中选择一个作为当前用户名，并将其发送给 LBS 服务器。在这里假设用户访问 LBS 是零星的，意味着两个连续的应用程序请求之间是有时间间隔的，所以攻击者很难将两个用户名联系成同一个用户，这种方法可以有效地抵制 LBS 服务器的长期统计攻击。

3　实验仿真与安全性分析

3.1　安全性分析

匿名算法中通过区域扩张后每个网格是真实用户位置的概率是相同的，这是隐私保护方法能够抵抗推理攻击的重要依据。使用 p_i 和 p_j 表示匿名区域中两个任意网格单元的位置 c_i 和 c_j 是真实用户的位置概率，通过本文算法的步骤，可以使匿名区域中的每个网格单元的查询概率基本相同，所以 $p_i = p_j$。但当匿名区域查询概率分布极度不均匀时，本文算法可能难以找到适合的匿名区域，即在这种情况下的区域中，攻击者能以大概率得出真实用户的位置。

另一方面对于基于长期统计的攻击模型，本文采取假名的方法，即在零星的查询过程当中，每一次的查询都使用不同的名字，所以即使攻击者有长期统计的数据也依然不能准确推断真实用户的位置。但同一假名使用次数越高越容易泄露用户的位置隐私，所以保证用户使用效率的情况下存储假名的个数有待确定。

3.2　实验仿真及分析

实验仿真将 GAGP 算法和其他三种算法在熵值度量、执行的成功率、生成匿名区域的面积及计算成本四方面作出比较。其中：GAGP 为本文提出的方法；Random 为不考虑查询概率，通过随机选择生成匿名区域的算法；GridDummy[17] 是在满足用户隐私要求的情况下获取匿名区域的方法；en-DLS[18]是基于从用户的历史位置提出的虚拟位置选择的算法。

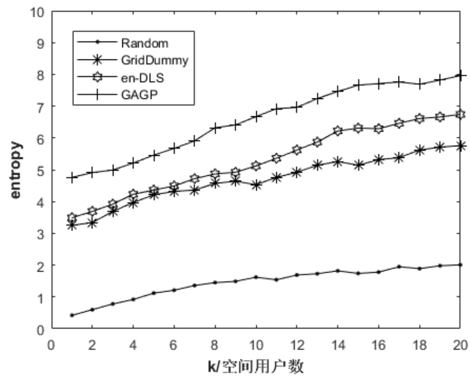


图 5　熵值和空间用户数的关系

Fig. 5　Relationship between entropy and the number of spatial users

图 5 表示将四种算法平均熵值以 k 值之间的关系进行比较的情况。可以看出随 k 的增加生成的匿名区域的平均熵值也会增加，因而使得用户真实位置所在网格单元的不确定性也就越大。因 GAGP 算法使用改进的网格划分方法将空间划分为查询频率基本相同的匿名空间，又使用了假名的方法，所以从结果可以看出 GAGP 算法的位置隐私保护在基于长期统计的攻击方式下效果较好。

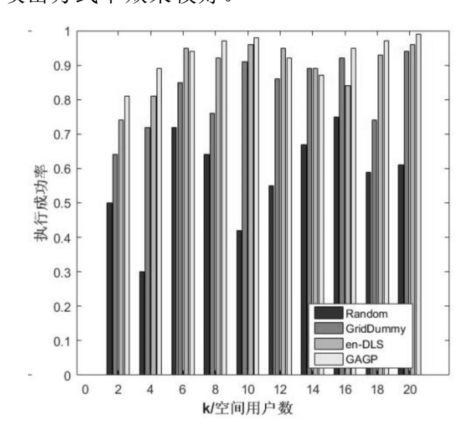


图 6　执行成功率和空间用户数的关系

Fig. 6　Relationship between the success rate of execution and the number of spatial users

图 6 表示使用本文及其他三种算法在以 LBS 为攻击者并且使用根据查询概率攻击的情况下，分析执行成功率的情况。由于 GAGP 算法是基于查询频率的区域扩张的方法，使得每个区域的查询频率基本相同，可以根据区域扩张将真实用户隐藏起来，所以攻击者很难根据概率找到真实用户。所以在应对基于查询频率的攻击时 GAGP 算法的隐藏效果更好。

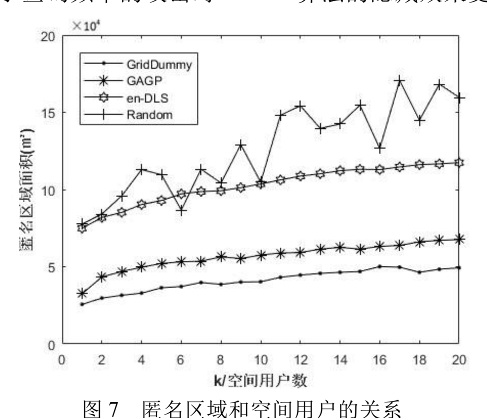


图 7　匿名区域和空间用户的关系

Fig. 7　Relationship between anonymous regions and spatial users

由于匿名区域的大小和用户位置隐私密切相关，所以图 7 表示 GAGP 算法和别的算法在空间用户数不断增多的情况下做生成匿名区域面积的比较。Random 因为是随机选择生成区域所以随着 k 值的增加不发生规律性变化，其他三种算法的变化都较为平缓。实验结果可以看出，随着空间用户数的增加，空间网格内总密度增大，匿名空间的代价也会相应减小，较小的空间内就能满足用户保护自己隐私信息的要求。所以虽然空间用户数较小时 GAGP 算法生成空间大于 GridDummyd 算法，但随着空间用户数的增加，匿名空间并没有过分增大，所以本文算法结与 en-DLS 和 Random 算法相比效果较好。

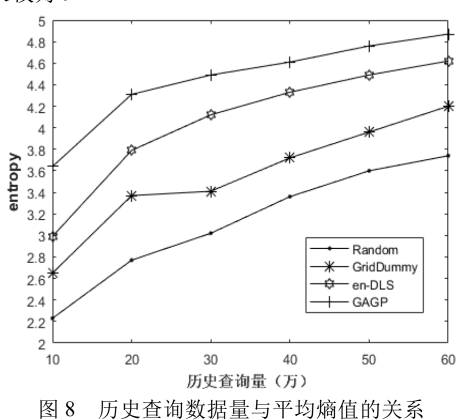


图 8　历史查询数据量与平均熵值的关系

Fig. 8　Relationship between the amount of historical query data and the average Entropy

图 8 表示了空间用户数 k=10 时，四种算法生成匿名区域的平均熵会随着历史查询数据量的变化而变化的情况。可以看出随着历史查询量的增大，算法的平均熵值也都在增大，且趋于平稳。因为本文使用假名，使攻击者收到的数据量越大，用户名就越多，从而基于长期统计的攻击模型就不易找出真实用户的隐私信息。所以 GAGP 算法的平均熵值明显大于其他三种算法，也可以说明本文算法效果较好。但到达一定数据量时，平均熵值不会再发生变动，且历史数据的增加也会带来额外的系统开销。

4　结束语

本文提出了一种基于遗传算法的空间网格划分的隐私保护方法。用户最初可以根据自己对隐私保护程度的偏好来设置初始匿名空间划分的粒度，然后第三方代理根据该地区的查询频率扩展网格，该方法使得重新划分后的空间网格查询频率基本相等，从而保证用户难以被攻击者识别。但当匿名区域查询概率分布极度不均匀是，GAGP 算法可能会出现无法找到最大阈值的网格单元即不能很好保护用户隐私的情况，另外 GAGP 算法主要针对的是用户在一个点不动的情况，用户连续移动位置情况的隐私保护方法将是未来研究工作的重点。

参考文献：

[1] 张磊，马春光，杨松涛，等．基于属性基加密的用户协作连续查询隐私保护策略 [J]．通信学报．2017, 38 (9): 76-85. (Zhang Lei, Ma Chunguang, Yang Songtao, *et al.* Privacy Protection Strategy of user Cooperative continuous query based on Attribute-based encryption [J]. Journal of Communications. 2017, 38 (9): 76-85.)

[2] Relish M, Young P, Vogel V, *et al.* Trusted third party for medication adherence [J]. Circulation: Cardiovascular Quality and Outcomes, 2016, 9 (S2): A266.

[3] Chen Jing, He Kun, Yuan Quan, *et al.* Blind filtering at third parties: an efficient privacy-preserving framework for location-based services [J]. IEEE Trans on Mobile Computing, 2018, 17 (3): 2524-2535.

[4] Zhang Shiwen, Liu Qin, Lin Yaping. Anonymizing popularity in online social networks with full utility [J]. Future Generation Computer Systems, 2017, 72 (7): 227-238.

[5] Ma Tinghuai, Jia Jing, Xue Yu, *et al.* Protection of location privacy for moving kNN queries in social networks [J]. Applied Soft Computing, 2018, 66 (3): 525-532.

[6] Roman Schlegel, Chi-Yin Chow, Duncan S. Wong, *et al.* User-Defined Privacy Grid System for Continuous Location-Based Services [J]. IEEE Trans on Mobile Computing, 2015, 14 (10): 2158-2172.

[7] 张磊，马春光，杨松涛，等．关联概率不可区分的位置隐私保护方法 [J]．通信学报．2017, 38 (8): 37-49. (Zhang Lei, Ma Chunguang, Yang Songtao, *et al.* An indistinguishable location privacy protection method based on association probability [J]. Journal of Communications. 2017, 38 (8): 37-49.)

[8] Niu Ben, Li Qinghua, Zhu Xiaoyan, *et al.* Enhancing privacy through caching in location-based services[C]//Proc of IEEE Conference on Computer Communications. Piscataway, NJ: IEEE Press, 2015: 1017-1025..

[9] Wang Shengling, Hu Qin, Sun Yunchuang, *et al.* Privacy preservation in location-based services [J]. IEEE Communications Magazine, 2018, 56 (3): 134-140.

[10] Sun Gang, Liao Dan, Li Hui, *et al.* L2P2: a location-label based approach for privacy preserving in LBS [J]. Future Generation Computer Systems, 2017, 74 (9): 375-384.

[11] Li Xinghua, Miao Meixia, Liu Hai, *et al.* An incentive mechanism for K-anonymity in LBS privacy protection based on credit mechanism [J]. Soft Computing, 2017, 21 (7): 3907-3917.

[12] Rohilla A, Khurana M, Singh L. Location privacy using homomorphic encryption over cloud [J]. International Journal of Computer Network and Information Security, 9(8):32-40

[13] Yin C, Sun R, Xi J. Location privacy protection based on improved k-value method in augmented reality on mobile devices [J]. Mobile Information Systems, 2017 (12): 1-7.

[14] Dargahi T, Ambrosin M, Conti M, *et al.* ABAKA: a novel

attribute-based *k*-anonymous collaborative solution for LBSs [J]. Computer Communications, 2016, 85 (7): 1-13.

[15] Ni Weiwei, Gu Mingzhu, Chen Xiao. Location privacy-preserving K nearest neighbor query under user's preference [J]. Knowledge-Based Systems, 2016, 103 (7): 19-27.

[16] Huang Yan, Cai Zhipeng, Bourgeois A G.. Search locations safely and accurately: a location privacy protection algorithm with accurate service [J]. Journal of Network and Computer Applications, 2018, 103 (1): 146-156.

[17] Sun Yanmin, Chen Min, Hu Long, *et al.* ASA: against statistical attacks for privacy-aware users in location based service [J]. Future Generation Computer Systems, 2017, 70 (5): 45-58.

[18] Niu Ben, Li Qinghua, Zhu Xiaoyan, et al. Achieving k-anonymity in privacy-aware location-based services [C]// Proc of IEEE Conference on Computer Communications. Piscataway, NJ: IEEE Press, 2014: 754-762.